# Dictionnaires électroniques et normalisation - éléments de réflexion

Laurent Romary

# Dictionnaires électroniques et normalisation - éléments de réflexion

Laurent Romary

avril 1997

## 1. Du support papier à l'information électronique.

L'informatisation d'un ouvrage tel que le Trésor de la Langue Française (TLF) doit pouvoir se situer dans la perspective plus large de la transformation de l'information textuelle — quelque soit son origine, littéraire, manuscrit, etc. — d'un support papier, par essence relativement pérenne dans sa forme et son contenu, à un représentation électronique pour laquelle ces simples notions de forme de contenu sont à la limite même difficiles à définir. Sans entrer dans une analyse fine de différents concepts qui accompagnent la notion même de livre électronique ou simplement de dictionnaire informatisé[1], essayons malgré tout de dégager quelques éléments de réflexion qui risquent de subordonner le succès de l'entreprise de numérisation d'un dictionnaire. Il serait en effet dangereux — surtout si l'on considère le volume de données que représente un dictionnaire de référence — de se lancer tête baissée dans des choix qui limiteraient l'utilité effective de l'entreprise de numérisation.

Considérons tout d'abord le support papier qui nous sert ici de référence[2]. Celui-ci contient de nombreuses informations de natures différentes qui résultent de choix liés à la fois au contenu, à la réalisation pratique de l'ouvrage, ou éventuellement à des considérations plus esthétiques. Les indications typographiques d'un dictionnaire permettent en particulier de séparer les différents champs contenus dans une entrée (indications grammaticales, définition, exemples, etc.). Leur sens est donc intimement lié à une *codification* préalable que des indications introductives, ou

[1] On trouvera notamment une réflexion très intéressante sur les usages possibles d'un dictionnaire informatique dans le rapport de D. Piotrowski et alii., *titre?????*.

[2] Une partie des difficultés que nous mentionnons ici n'ont déjà plus lieu d'être dans de nombreuses circonstances où le texte est directement produit sous forme informatisée, sans même parfois qu'une version sur support papier ne soit jamais produite (c'est le cas par exemple de certaines documentations techniques à usage interne à une entreprise). En anticipant sur un certain nombre de choix que nous suggérons, il apparait que de plus en plus d'éditeurs adopte directement des formats de représentation centrés sur le contenu, sur la base notamment de la norme SGML.

encore une pratique intertextuelle, permet de comprendre. Ces indications de forme typographiques sont d'ailleurs souvent associées à une certaine codification de la matière linguistique (abréviations pour les marques de catégorie grammaticales par exemple) qui en facilite encore plus la lecture.

De son côté, le support informatique impose ses propres contraintes qui nécessairement introduiront un écart par rapport à la version papier de référence. Il est ainsi clair qu'il serait vain d'attendre une lisibilité du contenu qui soit *équivalente* lorsque l'on passe du papier à l'ordinateur. De façon plus large, on peut immédiatement percevoir qu'en aucun cas on ne peut obtenir une fidélité parfaite par rapport au texte d'origine, quand bien même cette notion aurait un sens. Sans entrer dans ce débat, on peut s'interroger sur la façon dont il faut utiliser les informations contenues dans un dictionnaire papier. Prenons par exemple le cas d'un mot marqué en caractère gras dans une entrée de dictionnaire. Différentes solutions nous sont offertes pour le coder de façon électronique. On peut tout simplement occulter cette information et ne garder que la chaîne de caractère. On peut aussi - sous un format à déterminer - mémoriser l'indication typographique telle quelle (/le mot est en gras/). Enfin, on attacher au mot une indication *sémantique* correspondant à la signification du marqueur typographique initial (ex. /le mot est une entrée du dictionnaire/). On constate que ces différents choix dépendent intimement de ce que l'on souhaite privilégier lorsque l'on informatise un texte, de sa structure physique (son aspect extérieur) ou de sa structure logique (l'organisation de son contenu).

L'idéal est bien évidemment de déterminer un format de représentation qui préserve l'un et l'autre à la fois, ou au moins qui partant de l'un permette de reconstruire l'autre. L'expérience montre qu'il est bien plus facile de reconstituer l'apparence d'un texte à partir d'une représentation de sa forme logique, d'une feuille de style (ou équivalent) décrivant des équivalences entre les différents éléments repérés et la façon dont ceux-ci doivent être représentés, et enfin quelques règles typographiques. Comme nous ne reviendrons pas sur ces aspects, mentionnons que des normes telles que DSSSL[3] commence à apparaître pour unifier les descriptions de feuilles de styles pour des documents SGML. Clairement, adopter comme format privilégié quelque chose qui se rapproche plutôt de la structure logique d'un texte permet de

[3] Document Style Semantics and Specification Language (ISO IEC 10179, 1996)

s'affranchir du format effectif dans lequel on voudra produire une visualisation de celui-ci (HTML, postcript, etc.).

Une autre raison de privilégier un certain écart entre le texte tel qu'il apparaît sur support papier et une représentation électronique a trait aux nouveaux usages que l'on souhaite pouvoir associer au texte électroniques. En effet, la souplesse du support électronique, par exemple par la possibilité d'ajouter des informations qui ne seront pas immédiatement visibles à l'utilisateur (à des fins d'indexation par exemple), ou encore d'annoter dynamiquement les données proposées, permet d'imaginer bien d'autres accès qu'une recherche par mot vedette dans un dictionnaire ou d'autres visualisations que de présenter une entrée dans son intégralité. Dès lors, il faut envisager une représentation informatique d'un texte et a fortiori d'un dictionnaire comme une structure ouverte propre à devenir la matière première de nombreux usages différents, et non plus comme un produit fini, dans l'esprit de ce que peut être un ouvrage édité.

## 2. Pourquoi normaliser les données ?

Une fois posée la question du contenu informationnel que l'on désire mettre sur support électronique, l'étape suivante consiste à définir un format de représentation de ce contenu. Remarquons tout d'abord que le champ des possibles en la matière n'est pas infini de par la nature même des contraintes techniques inhérentes aux ordinateurs et aux logiciels que nous avons à notre disposition. De fait, il ne viendrait à l'idée de personne de réinventer une architecture complète d'ordinateur du simple fait que l'on doivent informatiser un dictionnaire. Cette remarque n'est pas totalement anodine puisque nous allons devoir évaluer jusqu'à quel point il nous est possible de faire des choix qui nous soient propres. Prenons par exemple la question du codage des caractères d'un texte. L'histoire de l'informatique, fortement ancrée sur le monde anglo-saxon, a pendant longtemps rendu difficile l'usage de caractères accentués tel que nous les connaissons en Europe de l'ouest. Au cours des années 70 et 80, chaque constructeur d'ordinateur ou fournisseur de logiciel destiné au travail sur une autre langue que l'anglais, a ainsi proposé un mode de représentation de ces caractères « exotiques » sans véritablement qu'il ait de concertations entre les différents acteurs concernés. Ainsi, les mondes du PC (sous MS-DOS), du Macintosh d'Apple, ou des stations de travail travaillant sous UNIX se sont-ils interdit tout échange transparent de données textuelles, opération qui nécessite systématiquement des filtres spécifiques de conversion. Devant l'ampleur des difficultés rencontrées, des groupes de

normalisation se sont alors penchés sur la définition d'un format de codage qui soit parfaitement identifiable et applicable à la plupart des types d'ordinateurs recensés. Le français, ainsi que les langues européennes de la façade ouest de notre continent, peuvent de la sorte être représentées à l'aide de la norme ISO 8859-1[4], adoptée progressivement par l'essentiel des constructeurs. Cette entreprise de normalisation ne s'est d'ailleurs pas arrêtée là puisque qu'une norme plus complète a récemment vu le jour, intégrant en un même ensemble, les représentations écrites des différentes langues mondiales (caractères romains, slaves, arabes, chinois etc.) au sein de ce qui est connu sous le nom d'UNICODE. De la sorte, nous disposons actuellement d'un support relativement stable pour représenter de façon uniforme toute matière textuelle sur un support électronique, sans qu'il soit nécessaire de se poser trop de questions.

Le niveau de normalisation atteint pour les informations de bas niveau tel que l'encodage des caractères n'a pas nécessairement son équivalent pour la représentation des contenus associés au texte ou au dictionnaire. On est ainsi en droit de se demander si l'on peut se permettre de s'appuyer sur des normes — disons plutôt des recommandations — qui émergent à peine au sein d'une communauté scientifique relativement hétérogène, à cheval entre informatique et sciences humaines. Remarquons tout d'abord que pour ce qui est du texte électronique, il existe un véritable débat sur la réelle nécessité d'ajouter des marques de structuration à un texte simplement représenté sous la forme d'une suite de caractères. Cependant, bien qu'une partie de la discussion sur la nécessité d'un certain niveau de normalisation s'applique tout à la fois au texte et au dictionnaire, ce débat préalable n'a pas lieu d'être pour ce dernier. En effet, un dictionnaire est par essence une matière structurée qu'il faudra, d'une manière ou d'une autre, représenter à l'aide d'un code spécifique. Il serait en effet inimaginable de gérer une version électronique d'un dictionnaire sans repérer au moins la séparation entre ses entrées, ainsi que les différents champs qui sont classiquement identifiés à l'aide de variations typographiques dans la version papier correspondante. Bien plus encore que pour le texte écrit, la typographie joue un rôle presque exclusivement sémantique pour le dictionnaire. Les éditeurs ont ainsi adoptés en

[4] Cette norme, connue aussi sous le nom d'ISO Latin possède la caractéristique de ne pas intégrer le caractère ligaturé œ. Comme toute norme, la question se pose alors la question de savoir si l'on « fait avec » (par exemple en utilisant une entité SGML - œ), ou si l'on se lance dans la définition d'une autre représentation…

interne différentes représentations, comme par exemple Harper-Collins pour son *Cobuilt Student's dictionnary*, dont nous reproduisons partiellement une entrée :

```
[HW] dab
[PR] /d*!ab/,
[IF] dabs, dabbing, dabbed
[LE]
[MB]
[MM] 1
[GR] VB [GS] with [GC] OBJ [GS] and [GC] ADJUNCT
[DT] If you [HH] dab [DC] a substance onto a surface, you put it there
with quick, light, strokes. If you [HH] dab [DC] a surface with
something, you touch it quickly and lightly with that thing.
[XB]
[XX] She dabbed some powder on her nose.
[XX] He dabbed the cuts with disinfectant.
[XE]
[ME]
[MB]
[MM] 2
[GR] COUNT N
[DT] A [HH] dab [DC] of something is a small amount of it that is put
onto a surface.
[XB]
[XX] She returned wearing a dab of rouge on each cheekbone.
[XE]
[ME]
[MB]
[MM] 3
[FB]
[GR] PHRASE
[DT] If you are a [HH] dab hand [DC] at something, you are good at doing
it;
[RN] an informal British use.
[AV] NBV
[FE]
[ME]
[EE]
```

Après une phase de compilation, la représentation physique associée à la représentation logique ci-dessus devient la suivant :

> **dab** /d*!ab/, **dabs, dabbing, dabbed. 1.** VB WITH OBJ AND ADJUNCT If you **dab** a substance onto a surface, you put it there with quick, light, strokes. If you **dab** a surface with something, you touch it quickly and lightly with that thing. *She dabbed some powder on her nose. He dabbed the cuts with disinfectant.* **2.** COUNT N A **dab** of something is a small amount of it that is put onto a surface. *She returned wearing a dab of rouge on each cheekbone.* **3.** PHRASE If you are a **dab hand** at something, you are good at doing it; an informal British use.

D'une certaine manière, dire qu'un dictionnaire électronique ne peut être formé que de la simple suite des caractères apparaissant dans sa version papier est un truisme. La difficulté est d'identifier les propriétés que l'on souhaite voir vérifiées par tel ou tel système de représentation. On peut ainsi, comme l'a proposé D. Piotrowski dans le rapport que nous avons déjà mentionné, adopter le schéma d'une base de donnée relationnelle, qui allie forte standardisation des entrées et fonctionnalités d'accès via un langage d'interrogation. Plus largement, nous pouvons essayer d'envisager un certain nombre de « bonnes » propriétés associées à un format de représentation.

Tout d'abord, le mode de représentation doit posséder un certain degré de lisibilité directe, c'est à dire indépendamment des outils éventuels qui vont servir à manipuler les données. Ce n'est pas tant que le dictionnaire électronique doive être lu comme le serait le support papier, mais l'opacité complète d'une représentation risque d'une part de ne pas permettre d'identifier facilement les données disponibles et d'autre part — ce n'est pas un argument à négliger — risque de décourager les utilisateurs potentiels de ces données qui auront l'impression que celles-ci leur seront devenues complètement inaccessible[5].

Un deuxième critère important est la souplesse du système de représentation vis à vis de la variabilité des données à représenter. Ainsi, une base de donnée relationnelle est particulièrement bien adapté à des objets associés à un nombre de paramètres stables et de taille relativement homogène. Les entrées d'un dictionnaire tel que le TLF sont particulièrement variables en taille, mais surtout en complexité, puisqu'il peur y avoir jusqu'à 5 ou 6 niveaux de récursivité dans l'emboîtement des rubriques.

La représentation doit aussi permettre l'ajout d'information supplémentaires (notes éditoriales, commentaires des auteurs) qui vont, comme nous l'avons signalé dans la première section, accompagner la matière du dictionnaire proprement dite. On le voit, un schéma trop rigide, qui collerait à la version papier de référence proscrirait un tel usage du support électronique.

Enfin, il faut probablement que le schéma de représentation choisi soit suffisamment générique pour qu'il puisse être appliqué — avec un minimum d'aménagement — à d'autres dictionnaires.

---

[5] On connait de nombreux cas d'utilisateurs de l'*Oxford Text Archive* qui dans un premier temps enlevaient les balises contenues dans certains textes qu'ils récupéraient de la base pour petit à petit les conserver par ce que disposant d'outils adéquates de manipulation. Le fait que la représentation restait « lisible » a permis de préserver le lien entre le serveur de ressources et l'utilisateur. L'idéal est bien évidemment, comme c'est le cas pour le serveur Silfide, que le mode de *présentation* des données puisse rendre transprent la *représentatio*n sous-jacente.

D'une part, il serait fastidieux de reprendre un travail complet de réflexion à la création de chaque nouvel ouvrage électronique, et d'autre part c'est une condition nécessaire à la réutilisabilité des outils créés pour manipuler ce type de données. La question qui se pose ici est en fait double ; d'un coté, il s'agit de déterminer en quoi un dictionnaire particulier est suffisamment spécifique pour justifier l'introduction de mécanismes *ad hoc* de représentation et de l'autre, il peut être bon de s'assurer que le mode de représentation propose des mécanismes de paramétrisation des structures adoptées et ne fixe donc pas de manière rigide un cadre qui là encore découragerait un futur concepteur de dictionnaire de l'utiliser.

La taille du *Trésor de la Langue Française* et l'entreprise que représente de fait son informatisation ne doit pas faire oublier qu'il s'agit d'une matière qui doit être rendue accessible au plus grand nombre, au sein d'un paysage culturel (cyber-culturel?) où de nombreux autres ouvrages, dictionnaires ou simples textes, auront été ou sont destinés à être eux-mêmes informatisés. La généricité de la représentation, parmi les différents critères que nous avons exposés, nous semble ainsi l'un des plus important. C'est l'assurance que le travail nous fournissons ne sera pas inutile, simplement parce qu'il serait trop lourd de le relier aux autres éléments du paysage.

## 3. Quelle normalisation ? - la TEI à l ′épreuve des dictionnaires.

L'ensemble des questions que nous nous sommes posées dans ce rapport ne sont guère originales si l'on considère l'important débat qui a secoué la communauté scientifique intéressée par la représentation de données linguistiques informatisées depuis 10 ans. En effet, en 1987, différentes associations savantes (notamment autour de *l'Association for Computers and the Humanities*), et des représentants de la plupart des équipes du domaine ont décidé de définir des recommandations pour la représentation et l'échange de textes électroniques. Très vite, il est apparu que la norme SGML qui émergeait à l'époque, était la seule réponse viable pour une telle entreprise, bien que n'étant pas nécessairement une solution idéale dans l'absolu[6]. Sur cette base, la *Text Encoding Initiative* a été fondée, et se trouve maintenant adoptée par de nombreux chercheurs et universitaires de par le monde (dans certains cas, le monde industriel tend à

[6] On repproche notamment à SGML son incapacités à représenter des structures multiples que pourraient associer à un même texte.

rejoindre le mouvement qui a été ainsi créé). Sous la forme de groupes de travail spécialisés, différents genres textuels ont été ainsi abordés, tels le roman, le théâtre, la transcription de matière orale, … et bien sur les dictionnaires.

Regardons rapidement les caractéristiques générales associées à la représentation des dictionnaires informatisés. Tout d'abord, il s'agit d'une description principalement arborescente des données reposant sur des *éléments* identifiés emboîtés les uns dans les autres. Une entrée de dictionnaire pourra ainsi avoir la forme suivante (extrait des directives de la TEI)[7] :

```
<entry>
  <form>
    <orth>r&eacute;moulade</orth>
    <pron>Remulad</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
    <gen>f</gen>
  </gramGrp>
  <trans>
    <tr>remoulade</tr>
    <tr>r&eacute;moulade</tr>
    <def>dressing containing mustard and herbs</def>
  </trans>
</entry>
```

où l'on distinguera des informations morphologiques (sous l'élément <form>), des informations grammaticales (<gramGrp>) et des informations correspondant à la traduction et à la définition dans la langue cible du mot (<trans>).

La description de telles entrées s'appuie sur une syntaxe générique du document correspondant, indiquant quelles peuvent être les possibilités d'occurrence de tel élément dans tel autre. Cette description du type de document ou DTD (*Document Type Definition* dans la norme SGML), donne une grande souplesse de représentation d'informations structurées, telles que celles contenues dans un dictionnaire.

A titre d'exemple, nous fournissons ci-dessous une possible représentation de l'entrée 'dab' du *Collin's Student Dictionnary* que nous avions présenté dans la section précédente :

```
<entry>
```

[7] Dans un document SGML, une balise telle que <entry> représente le début de l'élément, et </entry> en représente la fin. Des couples attribut=valeur, insérés dans la balise ouvrante d'un élément permettent d'apporter des informations supplémentaires au niveau de la représentation considérée (par exemple pour fournir un numéro d'ordre à une suite d'éléments de type <entry>). Les caractères accentués sont parfois codés à l'aide d'entités SGML (exemple é sera codé é). L'exemple est une transcription d'une entrée du *Collins Robert French-English English-French Dictionary* (Beryl T. Atkins et al., London: Collins, 1978, rpt. 1983).

```
        <form>
              <orth>dab</orth>
              <pro>/d*!ab/</pro>
        </form>
        <form type=infl>
              <orth>dabs</orth>
              <orth>dabbing</orth>
              <orth>dabbed</orth>
        </form>
        <gramGrp>
              <pos>subst. fém.</pos>
        </gramGrp>
        <sense n='1'>
              <usg type=gram>VB with OBJ and ADJUNCT</usg>
              <def>If you dab a substance onto a surface, you put it there
              with quick, light, strokes. If you dab a surface with
              something, you touch it quickly and lightly with that
              thing.</def>
              <eg>She dabbed some powder on her nose.</eg>
              <eg> He dabbed the cuts with disinfectant.</eg>
        </sense>
        <sense n='2'>
              <usg type=gram>COUNT N</usg>
              <def>A dab of something is a small amount of it that is put
              onto a surface.</def>
              <eg>  She returned wearing a dab of rouge on each
              cheekbone.</eg>
        </sense>
        <sense n='3'>
              <usg type=gram>PHRASE</usg>
              <def> If you are a dab hand at something, you are good at
              doing it; an informal British use.</def>
        </sense>
  </entry>
```

De fait, le TLF possède une structure beaucoup plus complexe que ce dictionnaire à l'usage d'étudiants anglicistes. L'annexe I de ce rapport présente par le détail une proposition de codage pour le TLF, illustré d'un exemple correspondant à l'essentiel de l'entrée 'quarantaine'. De nombreuses difficultés restent à résoudre, et nous nous y travaillons en collaboration avec nos collègues de l'INaLF. Il reste que les résultats obtenus sont relativement encourageant et laisse présager que le TLF informatisé pourra à moyen terme s'intégrer facilement dans le paysage électronique des sciences humaines.

Avant de conclure sur la normalisation des dictionnaires informatisés, il est peut être nécessaire de dire un mot sur la place que doit occuper une telle représentation dans la chaîne de traitement qui va de la production de l'information électronique, jusqu'à la présentation des données à un utilisateur final. Remarquons tout d'abord qu'il est nécessaire d'identifier, pour tout texte électronique, qu'il s'agisse d'un dictionnaire ou non, une version de référence sur laquelle vont

effectivement porter les différentes modifications que subit la matière électronique (ajouts, révisions etc.). La version de référence doit notamment contenir différentes informations documentaires (auteurs, responsables, caractéristiques du contenu) qui permettent de l'identifier parfaitement. De ce point de vue, la TEI, via la notion d'entête (*TeiHeader*), a prévu une telle documentation jusqu'à un niveau de détail qui couvre largement les besoins qui pourraient s'avérer nécessaire en la matière. De plus, nous avons signalé la souplesse qu'offrait SGML pour représenter des informations indépendamment de ce qui sera *in fine* présenté ou non à un utilisateur. Nous pensons donc que la représentation présentée plus haut a tout les atouts pour servir de version de référence à un système qui gérerait une chaîne d'édition ou au moins de gestion d'un dictionnaire électronique.

Cependant, si la taille du dictionnaire est relativement grande, il est clair qu'un fichier SGML risque d'être trop verbeux et mal adapté à un accès directe par le mot vedette par exemple. Dans un tel cas, il faudra envisager des représentations plus optimales (fichiers inverses etc.). Cependant, ces représentations spécifiques n'ont de raison d'être que pour des traitements bien particuliers et il est clair que différents traitements risquent de conduire à des représentations différentes. Par ailleurs, comme nous l'avons observé dans la première partie de ce texte, les formats de présentation de l'information peuvent être bien différents de ce qui apparaît dans un document SGML ; mais il existe de nombreux formats de ce type qui sont loin du niveau de normalisation que peut présenter SGML.

En substance, même si cela présente un certain nombre de difficultés, il nous semble qu'un document normalisé au format SGML suivant au plus près les directives de la TEI est la solution la plus raisonnable pour assurer une pérennité et surtout un bon usage à un dictionnaire que l'on souhaiterait mettre sous une forme informatisé, qu'il s'agisse de la rétroconversion d'un texte existant sur support papier, ou simplement de la création d'un nouvel ouvrage.

## 4. Perspectives - vers un accès en ligne au sein du serveur Silfide.

Il resterait un point essentiel à aborder, celui de l'interface offerte à un utilisateur qui voudrait accéder au TLF informatisé de façon souple et transparente. La perspective est ici d'intégrer l'ouvrage à un serveur de ressources linguistiques du type du serveur Silfide, défini dans le cadre conjoint du CNRS (Centre National de la Recherche Scientifique) et de l'Aupelf•Uref (Association des Universités Partiellement ou Entièrement de Langue Française). Ce serveur a

déjà démontré, dans le cas de textes informatisé, tout l'intérêt de mettre en ligne des textes électronique (codifiés suivant les directives de la TEI), sans imposer à l'utilisateur le maniement d'outils complexes que seul un expert informaticien pourrait mettre en œuvre. L'un de nos objectifs est maintenant de réaliser un travail similaire dans le domaine des dictionnaires informatisés et en particulier, de tester une telle interface sur la richesse offerte par le *Trésor de la Langue Française*. La description des résultats obtenus fera l'objet du prochain rapport…

## 5. Annexe I - Structure générale de l'entrée du TLF

Le TLF comporte une entrée principale non numérotée, elle même structurée en sous-partie correspondant à différentes valeurs. Il n'y a donc pas de suite d'homonymes qui justifierait l'usage de l'élément <superEntry> proposé par la TEI.

Le point d'entrée elle-même est composée de :

- le lemme, avec éventuellement quelques variantes
- une indication grammaticale (partie du discours)

Ainsi, le début de l'entrée *quarantaine* peut-elle être codée comme suit :

```
<entry>
      <form>
         <orth>QUARANTAINE</orth>,
      </form>
      <gramGrp>
         <pos>subst. fém.</pos>
      </gramGrp>
      …
</entry>
```

Les différentes sous-parties peuvent être facilement codées sous la forme de suite emboîtées de groupes « sémantiques » codée à l'aide de l'élément <sense>. On a ainsi pour *quarantaine* :

```
<entry>
      …
      <sense n='A'>
      </sense>
      <sense n='B'>
            <sense n='1'>
                  <sense n='a'>
                  </sense>
                  <sense n='b'>
                  </sense>
            </sense>
            <sense n='2'>
            </sense>
      </sense>
      <sense n='C'>
      </sense>
</entry>
```

De fait, la hiérarchie des entrées peut encore être plus imbriquées (liste à tirets, points etc.).

## Structure d'un bloc élémentaire

Nous considérons ici la structure d'un bloc élémentaire, tel que repéré par l'élément <sense>. Un tel bloc contient une suite d'informations comprenant en général, une définition, des exemples etc.

### *Représentation des définitions*

Les définitions représente une portion de texte décrivant le sens de l'entrée correspondante. Dans le TLF, une telle définition peut être soit originale, soit empruntée à une source extérieur. Dans le premier cas, on marque simplement le texte à l'aide de l'élément <def> :

```
<def>Ensemble, nombre de quarante, d'environ quarante. </def>
```

Dans le deuxième cas, on indique qu'il s'agit d'une citation (élément <cit>) en repérant le texte cité (<q>), ainsi que la source bibliographique (élément structuré <bibl>). Par exemple (cycloptère…) :

```
<def><cit>
      <q>,,Poisson osseux des mers froides, au corps très épais,
      globuleux, muni d'une ventouse formée par les nageoires
      ventrales``</q>
      <bibl>(<title>Lar. encyclop.</title>)</bibl>.
</cit>
</def>
```

### *Représentation des exemples*

Les exemples, marqués en italique dans le texte, peuvent être repérés à l'aide de l'élément <eg>. Qand l'exemple est issu d'une source citée dans le texte, on utilise l'élément <cite>, comme dans le cas des définitions.

## Indications diverses dans une entrée

Il s'agit d'indication précédent en général les définition et apportant une restriction particulière. Le codage Tei préconise la balise <usg> qui permet une indication plus fine du type de restriction.

### *Restrictions géographique*

Les indications de régionalismes (*Région,* etc.) sont indiquées comme suit :

```
<usg type=geo>…</usg>
```

### *Restrictions temporelles*

Les indication recueillies portent sur le caractère vielli de l'usge lexical répertorié. Le TLF utilise toute une série de marqueurs : vieilli, Vieilli, Vx., vx. etc. Le codage adopté est le suivant :

```
<usg type=time> vielli. </usg>
```

### *Restrictions grammaticales*

Deux possibilités s'offraient à nous :

```
<usg type=gram>…</usg>
```

où

```
<gramGrp>
      <pos>…</pos>
<gramGrp>
```

La deuxième solution semble meilleure car elle permet de gérer de façon uniforme toute les indications grammaticale (cf. entrées du dico). Nous l'avons utilisée pour toutes les indications reprenant une partie des catégorie grammaticales de l'entrée titre.

La première solution a été utilisée pour les restriction d'usage plus particulières. Par exemple : *Au plur., Au passif,* etc.

### *Fréquences*

Pour toutes les indications de type *rare, hapax* etc. la valeur d'attibut (type=freq) a été introduite. Celle-ci n'est pas référencée dans le chapitre 12 de la TEI, mais devrait probablement être ajoutée.

### *Indications stylistiques*

```
<ind>Au fig.</ind>
```

devient :

```
<usg type=style>Au fig.</usg>
```

De la même façon :

*péj.*

Les entrées complexes telles que :

*Au fig., fam.*

sont éclatées en plusieurs éléments <usg>

### *Indications de registre*

Toutes les indications du type : *fam., pop., Famillier, Trivial*

sont transcrites à l'aide de :

```
<usg type=reg>fam..</usg>
```

### *Indications grammaticales particulières*

Il s'agit d'indications d'emploi qui ne relèvent pas d'un gramGrp, mais qui relève néanmoins du domaine de la grammaire du français. On a ainsi :

*En apposition,*

## 6. Annexe II - L'entrée 'quarantaine' du TLF

```
<entry>
      <form>
         <orth>QUARANTAINE</orth>,
      </form>
<gramGrp>
   <pos>subst. fém.</pos>
</gramGrp><lb>
<sense n=A>A. _
      <def>Ensemble, nombre de quarante, d'environ quarante. </def>
      <eg><q>Quarantaine d'années, de kilomètres.  </q></eg>
      <eg><cit>
      <q>Le nombre des corps simples actuellement connus paraît être
      voisin de 72, parmi lesquels près de 30 n'ont qu'une importance
      tout à fait secondaire, ce qui réduit à une quarantaine le nombre
      des éléments réellement répandus dans la nature</q>
      <bibl>     (<author>Lapparent</author>,     <title>Minér.</title>,
      <date>1899</date>, <biblScope>p. 4</biblScope>)</bibl>.
      </cit>
      </eg>
      <eg><cit>
      <q>Sur cent trente hectares il y en avait une quarantaine de
      cultivables: dix en vignes, trente en céréales</q>
      <bibl>    (<author>Bosco</author>,    <title>Mas    Théot.</title>,
      <date>1945</date>, <biblScope>p. 248</biblScope>)</bibl>.
      </cit>
      </eg>
</sense><lb>
<sense n=B>B. _
      <I>
      <ind>En partic.</ind>
      <lb>
      <sense> 1.
            <def>Période  de  quarante  ou  d'environ  quarante  jours
            consécutifs.</def><lb>
            <sense> a)
                  <usg type=dom>RELIG. CATH., </usg>
                  <usg type=time>vieilli.  </usg>
                  <eg><cit>
                  <q>Innocent IV (...) accorda un an et quarante jours
                  d'indulgence (...). Sixte IV (...) cinquante années et
                  autant de quarantaines d'indulgence à tous les fidèles
                  (...) qui visiteraient les églises de l'ordre de saint
                  François</q>
```

```
      <bibl>    (<author>Montalembert</author>,    <title>Ste
      élisabeth</title>,   <date>1836</date>,   <biblScope>p.
      321</biblScope>)</bibl>.
      </cit>
      </eg>
      <lb>
      <sense> .
      <eg><q>(Sainte) quarantaine. </q></eg>
      <def>Carême. </def>
      <eg><cit>
      <q>Jeûner la quarantaine</q>
      <bibl> (<pub><I>Ac.<R></pub>)</bibl>.
      </cit>
      </eg>
      <eg><cit>
      <q>Ce temps de la sainte quarantaine était, au point
      de vue liturgique, admirable; la tristesse y allait
      grandissant chaque jour, avant que d'éclater (...) en
      les douloureux sanglots de la Semaine Sainte</q>
      <bibl>                    (<author>Huysmans</author>,
      <title>Oblat<ct><R>,         t.         2</ct></title>,
      <dat>1903</dat>, <biblScope>p. 42</biblScope>)</bibl>.
      </cit>
      </eg>
      <eg><cit>
      <q>Ces fêtes rustiques (...) par lesquelles les (...)
      chrétiens rompaient la pénitence de la sainte
      quarantaine</q>
      <bibl>       (<author>France</author>,       <title>J.
      d'Arc<ct><R>,   t.   1</ct></title>,   <dat>1908</dat>,
      <biblScope>p. 179</biblScope>)</bibl>.
      </cit>
      </eg>
      </sense><lb>
      <sense> .
      <usg type=geo>[En Lorraine]  <I></usg>
      <eg><q>Messe de quarantaine. </q></eg>
      <def>Messe célébrée à l'intention d'un mort (environ)
      quarante jours après son décès. </def>
      <eg><cit>
      <q>Le jour du «service de quarantaine», qui réunit à
      l'église tous les parents et voisins du mort
      (Lorraine)</q>
      <bibl> (<author>Menon, Lecotté</author>, <title>Vill.
      Fr.<ct><R>,      2</ct></title>,      <dat>1954</dat>,
      <biblScope>p. 102</biblScope>)</bibl>.
      </cit>
      </eg>
      <exe n=E><I>Une messe de quarantaine sera célébrée le
      dimanche onze novembre (...) à la mémoire de
      Monsieur<R> [<I>X<R>] <I>(...) décédé le 3 octobre
      1984<R> (<pub><I>L'Est Républicain</pub><dat><R>, 9
      nov. 1984</dat>, <biblScope>p. 4</biblScope>).</exe>
      </sense><lb>
</sense>
<sense> b)
      <def>Isolement (de quarante jours à l'origine) imposé
      aux personnes, aux animaux ou aux choses atteints ou
      contaminés par une maladie contagieuse ou susceptibles
      de l'être. </def>
      <eg><cit>
```

<q>Dans les régions où la typho-anémie sévit, il ne faut jamais introduire un nouveau cheval dans les écuries qu'après une quarantaine rigoureuse d'au moins un mois</q>
<bibl> (<author>Garcin</author>, <title>Guide vétér.</title>, <date>1944</date>, <biblScope>p. 226</biblScope>)</bibl>.
</cit>
</eg>
<lb>
<sense> .
<eg><q>Mettre en quarantaine. </q></eg>
<exe n=E><I>Son second fils a la rougeole: la maison est mise en quarantaine pour ne pas propager l'épidémie<R> (<author>Estaunié</author>, <title>Empreinte,</title> <date>1896</date>, <biblScope>p. 308</biblScope>):</exe>
<lb>
<exe n=D>1.Comme un matelot du bord était mort d'une sale maladie, les autorités ont cru que peut-être c'était la peste, et on nous a mis en <G>quarantaine<R>.
<author>Pagnol</author>, <title>Fanny</title>, <date>1932</date><biblScope>, I, 1<abbr type=superscription>er</abbr> tabl., 14, p. 54</biblScope>.</exe>
<lb><lb>
<eg><q> <G>SYNT. <I>Quarantaine sanitaire; formalités de quarantaine; hisser le pavillon de quarantaine; faire quelques jours de quarantaine; malades, voyageurs en quarantaine; maintenir, placer, retenir en quarantaine; soumettre des marchandises, un navire, des voyageurs à une quarantaine<R>.</q></eg></sense><lb>
<sense> _
<I>
<usg type=style>P. méton.</usg>,
<usg type=dom>MAR. </usg><I>
<eg><q>Pavillon de quarantaine. </q></eg>
<xr><lbl>V. </lbl><ref>pavillon ex. 4.</ref></xr>

<lb>
</sense>
<sense> _
<I>
<usg type=style>Au fig.</usg>
<R>
<def>Exclusion d'un groupe social. </def>
<eg><q>Tenir qqn en quarantaine. </q></eg>
<eg><cit>
<q>La quarantaine qu'on fait ainsi subir aux talents nouveaux, avant de les accepter et de les louer, cause des impatiences, comme toutes les quarantaines</q>
<bibl> (<aut><P>Sainte<R>-<P>Beuve</aut>, <title>Portr. contemp.<ct><R>, t. 3</ct></title>, <dat>1842</dat>, <biblScope>p. 308</biblScope>)</bibl>.
</cit>
</eg>
<exe n=E><I>Au Bechouanaland, l'homme qui est en quarantaine pour (...)<R> [<I>avoir tué un ennemi<R>]

<I>ne doit toucher personne, et son ombre ne doit pas effleurer ses enfants<R> (<pub><I>Jeux et sports</pub>, <date>1967</date>, <biblScope>p. 773</biblScope>).</exe>
<lb>
<sense> .
<eg><q>Mettre en quarantaine:</q></eg>
<lb>
<exe n=D>2.Grange (...) s'emballa. On se quitta bêtement. Et voilà ses fabriques en interdit: aucun compagnon n'aurait plus osé travailler pour lui; il aurait été <I>mis en <G>quarantaine<R> comme renégat.
<aut> <P>Pourrat</aut>, <title>Gaspard</title>, <date>1930</date>, <biblScope>p. 35</biblScope>.</exe>
<lb></sense><lb>
</sense>
</sense>
</sense>
<sense> 2.
<def>Âge de quarante ans, d'environ quarante ans. </def>
<eg><q>Atteindre, avoir, dépasser la quarantaine. </q></eg>
<eg><cit>
<q>Voilà la quarantaine qui approche; j'ai eu 37 ans le 12 décembre dernier</q>
<bibl> (<author>Flaub.</author>, <title>Corresp.</title>, <date>1859</date>, <biblScope>p. 307</biblScope>)</bibl>.
</cit>
</eg>
<eg><cit>
<q>Le trouble d'accommodation habituel, vers la quarantaine, est la presbytie</q>
<bibl> (<author>Macaigne</author>, <title>Précis hyg.</title>, <date>1911</date>, <biblScope>p. 283</biblScope>)</bibl>.
</cit>
</eg>
</sense><lb>
</sense>
<sense n=C>C. _
<xr type=syn><lbl>Synon. rare de </lbl><ref rend=italics>giroflée</ref>. </xr>
<eg><cit>
<q>Des quarantaines blanches fleurissaient les corbeilles</q>
<bibl> (<author>Zola</author>, <title>Page amour</title>, <date>1878</date>, <biblScope>p. 879</biblScope>)</bibl>.
</cit>
</eg>
<eg><cit>
<q>Quarantaines (...), Puissant encens du mois d'août!</q>
<bibl> (<author>Claudel</author>, <title>Poés. div.</title>, <date>1952</date>, <biblScope>p. 871</biblScope>)</bibl>.
</cit>
</eg>
</sense><lb>
<lb><lb></entry>